Developing a CLARIN compatible AAI solution for academic and restricted resources

Tommi A Pirinen **Daniel Jettka** **Hanna Hedeland**
Hamburger Zentrum für Sprachkorpora
Universität Hamburg, Germany
first.last@uni-hamburg.de

Abstract

In this article we introduce an authentication and authorisation infrastructure for a CLARIN-compatible digital repository. Due to data protection, the kind of corpora hosted by the repository normally cannot be made available under free licenses (e.g. Creative Commons), and only very few are available for general academic use. Most often, they are restricted to academic, non-commercial use and are only made available upon personal request. These characteristics were translated into general system requirements which resulted in the clear conceptual and institutional separation of authentication (single sign-on via Shibboleth) and authorisation (role-based rights management). Although implementations with similar ideas in mind exist in a number of different repository systems, we have found a comprehensive and seamless method to integrate it directly into our exisiting repository system based on Fedora Commons and Drupal/Islandora.

1 Introduction

In this article we describe a *AAI* (Authentication and Authorisation Infrastructure) solution for a digital repository at a research data centre for language corpora. Since the focus of the centre lies on various types of spoken language corpora, the resources hosted in the repository come with varying distribution types and levels of licensing, and only few of them are publicly available under e.g. Creative Commons licenses, while most are distributed as *restricted* (RES) resources.

The AAI solution is part of a digital repository architecture that has been developed since 2011 and is in production use since 2013. Due to a close connection to the CLARIN-D infrastructure and after thorough evaluation of other systems, the repository was built with the software Fedora Commons[1] and the Drupal[2] based Islandora framework[3]. This combination fulfills a number of basic needs: For instance, it guarantees direct compatibility with most CLARIN-D centres (seven of eight also use Fedora Commons), facilitating comprehensive data transfer between CLARIN-D centres if this should become necessary. Furthermore, Islandora provides several modules which proved very useful for our needs, such as a web-based GUI for fine-grained role management and adaptable resource-specific ingest and dissemination methods.

Although Fedora Commons and Islandora comprise the basis for our repository, the presented AAI solution can be seen as a general concept that could be adopted by other repository systems. While in our repository we started with a Shibboleth-based *SSO* (Single Sign-On) solution in addition to traditional password-based authentication (i.e. the Drupal AAI system), we are now moving towards outsourcing the authentication of all users to Shibboleth, while still using automatically and manually set roles controlling the access to protected resources.

In summary, the goal of this article is to discuss advantages and potential disadvantages of a conceptual and institutional separation of user authentication and authorisation, rather than technical details of the implementation. Apparently, our repository management workload can be reduced by the approach to a noticeable degree, in that we do not have to deal with the registration of users and potentially attached

[1] http://fedora-commons.org/
[2] https://www.drupal.org/
[3] https://islandora.ca/

communication cycles. For our users the approach leads to a higher degree of usability, since they do not need yet another account and password for accessing restricted data at our repository. Another profit is the enhanced level of security since passwords stay at the user's home institution (cf. CLARIN-D and DARIAH-DE call for action on federated identity[4]).

The rest of the article is organised as follows: Section 2 describes the relevant requirements for the AAI solution in more detail. In Section 3 we describe our authentication approach and in Section 4 the authorisation approach, while finally in Section 5 we summarise the lessons learnt from building the system and lay out possible future developments.

2 Requirements

The basic requirements of our AAI architecture arise from the circumstance that the vast majority of resources in the repository are classified as restricted (RES) resources, which means access can only be granted to identified users who provide a valid purpose and agree to the respective Terms of Use. Recent developments to facilitate data sharing for the academic community in general using federated login are therefore only applicable with additional features implemented to meet these requirements.

The corpus resources were originally distributed via a web server with limited means for advanced user management. However, this legacy solution allowed for flexible individual access for any approved user and use case, even if the user did not belong to the academic community in a narrow sense but wanted to use the resources for academic and non-commercial purposes. For example, a resource on community interpreting in hospitals could be useful in developing training materials for medical staff. For some resources, the user is required to sign a user agreement and send it to the centre in digital or analogue format. Considering this variation in restrictive licensing conditions, a basic system requirement arises from the fact that human resources for user management are limited and workflows for granting access should be as transparent and efficient as possible.

Some technical implications also arise from external sources, e.g. CLARIN's requirements[5] and the *Data Seal of Approval* (DSA) requirements[6]. CLARIN B type centres must support single sign-on via Shibboleth, which is of course relevant beyond CLARIN. The SSO solution should also be robust enough to support logins from universities that release limited or partial information about their users. This is a practical requirement that arises from common practices.

While many repositories do not need fine-grained authorisation mechanisms, e.g., because they only provide publicly available and freely usable resources and/or resources which are directly available upon successful login via SSO, there are several repositories in CLARIN with similar approaches and use cases, e.g. in CLARIN-DK[7] (Offersgaard et al., 2013), and at the Meertens' institute[8] (Windhouwer et al., 2016). CLARIN-DK uses eSciDoc (Offersgaard et al., 2011) that also makes use of Shibboleth as well as role-based authorisation,[9] and the Meertens institute uses a solution developed for TLA and their repositories: FLAT[10], which is also technically similar to our solution. Other solutions are developed for other repository software, e.g. within CLARIN DSpace[11] (Mišutka et al., 2015). Another repository system independent solution is the Resource Entitlement Management System (REMS (Linden et al., 2013)), which is in use for instance at the Language Bank of Finland[12], but was not directly usable in our case since it is intended to be used as a service, or alternatively seemed to require a high degree of implementation work.[13]

All in all, there are a number of different solutions and approaches to AAI for digital repositories which are useful in a variety of contexts. However, we could not identify a system that was as powerful

[4] https://www.clarin.eu/sites/default/files/clarin_dariah_call-for-action-aai.pdf
[5] http://hdl.handle.net/11372/DOC-77
[6] https://www.datasealofapproval.org/en/information/requirements/
[7] https://clarin.dk/clarindk/forside.jsp
[8] http://www.meertens.knaw.nl/cms/en/
[9] https://www.escidoc.org/JSPWiki/en/Security
[10] https://github.com/TLA-FLAT/FLAT
[11] https://github.com/ufal/clarin-dspace
[12] https://kielipankki.fi/
[13] https://confluence.csc.fi/display/REMS/Installation

and usable with respect to our needs as the combination of Drupal (Islandora) and XACML policies in Fedora Commons, which allow for the restriction of individual access (API-A) and management actions (API-M) for collections and single resources in the repository.

3 Authentication via Shibboleth

This section deals with the authentication part of our AAI approach, for which we have tested a number of solutions. In this article we focus our solutions based on Drupal. The initial solution was to use both Drupal's native user management and an additional Shibboleth module[14]. Both systems both have pros and cons, for example the use of Drupal authentication is easy to set-up and lets anyone access the system, however, given that most resources are rather restricted in usage terms, the amount of manual checking needed before granting access rights is higher. With the SSO solution, it is theoretically very easy to verify the user's current academic status and affiliation, however, this only works as long as the *IdP* (Identity Provider) provides the necessary information to authenticate users. Unfortunately, this is not the case for all institutions, which increases the support workload as well. For example, a number of universities that support SSO do not release any information about the user upon login[15]. However, in order to log the user into an account on Drupal, some form of permanent uniquely identifiable information would be necessary. We need to support at least a number of scenarios where end-users come from an institution with an IdP setup that doesn't release enough attributes[16], that come from outside the federation, or even outside academia. The solution we are moving towards now, is based on outsourcing authentication. Since a great number of institutions do not release sufficient attributes, in effect this means we have to rely on (academic and non-academic) CLARIN accounts and for this, we need to trust the CLARIN IdP and its entitlement settings for individual users, bearing in mind that the policies for user management of the CLARIN IdP cannot be compared to those of IdPs in regular national identity federations.[17]

4 Role-based Authorisation

The authorisation system of our repository has to reflect the circumstance that most of the existing resources come with specific access restrictions due to data protection. Apart from basic metadata, many parts of the resources contain information that can only be disseminated to identified persons upon request if they agree to specific terms of use. The process of granting access to a resource is explicitly documented in agreements with the data owners who may have posed additional requirements for the usage of the data. In some cases only certain parts of resources (e.g. only transcriptions, but no recordings) can be given to users, which means that the authorisation system has to be able to restrict and grant access on global (complete resource) and local (e.g. collection or file-based) level.

The authorisation for the *open* (PUB) resources and many *academic* (ACA) resources can be determined by the basic account information provided by the Shibboleth login identifying a user as being academic, i.e. either the user has a member-type affiliation to an academic organisation, or an academic-type entitlement, the latter being mainly the case for CLARIN user accounts. All of our users have to accept basic terms of service[18] once upon first login. Practically, this is the level of granularity that is necessary for many repositories' authorisation needs. For the restricted RES corpora the authorisation is handled on individual basis using a access request form that is presented to the user on the web page of the respective corpus.

In principle, every authorisation system for fine-grained user role management can be used to facilitate our AAI approach. In our repository, however, the access to resources is managed by Drupal's role system in conjunction with Fedora's XACML policies (Anderson et al., 2005)[19] for authorizing access to digital

[14] https://www.drupal.org/project/shib_auth

[15] https://lindat.mff.cuni.cz/services/aaggreg/

[16] https://www.clarin.eu/sites/default/files/clarin_dariah_call-for-action-aai.pdf, compare to CLARIN D-SPACE explanation of same issuehttps://github.com/ufal/clarin-dspace/wiki/Shibboleth-accounts---email,-netid-and-idp

[17] https://www.clarin.eu/content/clarin-identity-provider

[18] https://corpora.uni-hamburg.de/hzsk/en/corpus-enquiries-licenses

[19] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

objects and datastreams. The XACML policies allow for a fine-grained definition of possible actions (for object access and management) and an unrestricted number of roles bound to these actions. We connect our resources to a cascade of resource-specific role names which makes it possible to grant access on a global, intermediate, or local level. For instance, a transcript from a corpus could be accessible to users with the roles *demo-full*, *demo-transcripts*, and *demo-transcript-001*, so that we could give users exclusive access to the complete demo corpus, the transcripts of the corpus, or the single transcript only. The presence of these role names in the XACML policies does not necessarily mean that all of them have to be instantiated in the Drupal system. In contrast, they can be created in the Drupal backend and connected to users just when they are needed, so that the number of roles actually in use (around 50) differs greatly from the thousands of role names defined in the XACML policies.

The resource-specific roles are normally assigned to users manually (by any user with sufficient rights for the Islandora administration interface) after a user has sent a corpus access request that has then been approved by us and/or the data owner. In addition, some roles can be assigned to users automatically based on Shibboleth attribute matching. This makes it possible to grant direct access to (ACA) resources that can be made available to all academic users. The Drupal Shibboleth module we use for this task allows for the definition of rules, for instance taking into account the Shibboleth attributes *Shib-Identity-Provider* and *entitlement* to automatically decide which roles are assigned to a user upon login. This way users with academic entitlement can instantly access the ACA resources, while other users (with potentially commercial background) can sign in via Shibboleth but would still have to undergo the corpus request procedure even for ACA resources, which gives us the chance to personally evaluate the corpus request and the intended purpose of use.

5 Future Directions

In this article we have presented the basic concepts of an authentication and authorisation infrastructure for a CLARIN-compatible digital repository, focussing on the clear conceptual and institutional separation of authentication (single sign-on via Shibboleth) and authorisation (role-based rights management). Although the concept can and has been implemented in various ways in a number of repository systems, we have found a comprehensive and seamless method to integrate it directly into our exisiting repository system based on Fedora Commons and Drupal/Islandora.

While the system described is fully operational and working, there are still future improvements to be made and certain issues to be resolved. One of the issues arises from related resources/services hosted outside of Drupal and Fedora, such as integrated web applications. While it is possible to integrate Apache's Shibboleth module to restrict access to services to authenticated or academic users as defined by Shibboleth attributes, the authorisation against roles defined in the rights management system of Drupal requires a non-trivial amount of modifications to the web application or resource. A solution might be the integration of a more powerful external AAI module, though the increased complexity might come with a high cost if the maintenance cannot be guaranteed.

Another problem is that IdPs do not always release sufficient information (Shibboleth attributes) about their users, which leads to a suboptimal user experience. In the future, this can possibly be alleviated with e.g. new additions to Shibboleth like uApprove[20], a user consent module that lets users customise the attributes released, or similar built-in features in newer versions of the IdP software, especially with an increasing number of users of the CLARIN infrastructure asking their institutions to allow them to release attributes when required.

References

[Anderson et al.2005] Anne Anderson, Anthony Nadalin, B Parducci, D Engovatov, H Lockhart, M Kudo, P Humenn, S Godik, S Anderson, S Crocker, et al. 2005. extensible access control markup language (XACML) version 2.0. *OASIS*.

[20] `https://www.switch.ch/aai/support/tools/uapprove/`

[Linden et al.2013] Mikael Linden, Tommi Nyrönen, and Ilkka Lappalainen. 2013. Resource entitlement management system.

[Mišutka et al.2015] Jozef Mišutka, Amir Kamran, Ondřej Košarko, Michal Josífko, Loganathan Ramasamy, Pavel Straňák, and Jan Hajič. 2015. Linguistic digital repository based on DSpace 5.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

[Offersgaard et al.2011] Lene Offersgaard, Bart Jongejan, and Bente Maegaard. 2011. How danish users tried to answer the unaskable during implementation of clarin.dk. In *SDH 2011-Supporting Digital Humanities*.

[Offersgaard et al.2013] Lene Offersgaard, Bart Jongejan, Mitchell Seaton, and Dorte Haltrup Hansen. 2013. CLARIN-DK - status and challenges. In *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 20*, number 89, pages 21–32. Linköping University Electronic Press; Linköpings universitet.

[Windhouwer et al.2016] Menzo Windhouwer, Marc Kemps-Snijders, Paul Trilsbeek, André Moreira, Bas Van der Veen, Guilherme Silva, and Daniel Von Rhein. 2016. FLAT: constructing a CLARIN compatible home for language resources. In *LREC 2016: 10th International Conference on Language Resources and Evalution*, pages 2478–2483. European Language Resources Association (ELRA).